# NLP2 Project A

Trustworthy Bias Measures for Language Models and Word Embeddings

TA: Oskar

## 1 Introduction

The field of NLP has seen great success with the development of language models (LMs) based on deep neural networks, which leverage an increasing amount of training data and computing power to learn key features of natural language. An unfortunate byproduct of this paradigm, is the added complexity of state-of-the-art LMs, which have become notoriously opaque—we often refer to these models of billions of parameters as “black boxes” as we don’t exactly know what mechanisms and representations are learned. While we know that these LMs are prone to learning and amplifying social biases from e.g. the training data, researchers still lack the proper tools for measuring these biases in NLP systems reliably. Even though many bias measures have been proposed in the literature, it is unclear which of these are actually trustworthy [1, 2, 8, 7].

Not only the “black-box” nature of LMs makes measuring social biases difficult. Researchers also lack gold-standard labels for how biased a model is, so we cannot easily calibrate and test our bias measures. One could even argue that it is impossible to have one definite ground-truth, as social biases are inherently subjective and context-dependent [1, 14]. Considering these challenges, it is clear that we have to be careful in how we design and test a bias measure.

Two useful concepts for describing the trustworthiness of bias measures that can help in this task are: i) reliability (*is your measure consistent?*) and ii) validity (*are you measuring what you intend to measure?*) [9, 15, 4].

## 2 Assignment

In this project, you will familiarize yourself with some techniques for measuring bias in NLP models as well as certain strategies for testing their validity and reliability.

For the first week, there is a tutorial with some assignments (5% of your grade) that help you get started with some techniques for measuring bias. After doing these assignments, you will start with your own experiment where you choose to test the validity and reliability of one of the following bias measures:

1. WEAT [5];
2. Bias Direction (e.g., [3]; see see e.g. this AllenNLP guide); and
3. CrowS-Pairs [10, 11].

The first two are for static word embeddings (e.g., word2vec, glove, fasttext), while CrowS-Pairs is designed for a language model (originally BERT, but it can be adapted to suit e.g. GPT-2 when using perplexity to compare sentences).

**You will implement this bias measure (if applicable) and then assess its validity and reliability**. Note that some bias measures are harder to implement than others, but we take into account the complexity of the *whole* experimental setup.

More information on the concepts of validity and reliability can be found in Sections 3 and 4 of [15].

## 3 Deliverables

1. PDF with assignments from the tutorial, due April 28, 2023 at 23:59

2. Jupyter notebook, due April 28, 2023 at 23:59. The notebook should contain the entire pipeline from data generation to model training to the analysis conducted. Functions or classes are allowed to be defined in Python files externally, as long as the main functionality is listed in the notebook. We recommend training your models on GPUs through the Google Colab service.[1]

3. Short paper, due April 28, 2023 at 23:59. The short paper should use the ACL conferences template[2] and contain four pages (references excluded). A suggested page distribution is as follows:

    (a) **Abstract:** summarise the research in a short piece of text that emphasises your contributions and findings (0.1 pages);

    (b) **Introduction:** introduce the reader to your research area, summarise your contributions and highlight the relevance of your research, provide a clear and explicit problem statement as well as your research questions (0.5 pages);

    (c) **Background** summarise research papers relevant for your work. Be brief, since this is a short paper (0.4 pages);

    (d) **Approach:** dependent on the particular project, this section should detail the tasks or models designed (1 page);

    (e) **Experiments and Results:** detail the precise experimental setup used and the results of your evaluation measures (1 page);

    (f) **Discussion:** interpret the results of analysing the bias measures. You should also give suggestions for future work. (1 page).

## 4 Recommended reading

1. Gonen and Goldberg [8]

2. Orgad and Belinkov [12]

[1]Visit Google Colab: `https://colab.research.google.com/`

[2]Visit the template: `https://github.com/acl-org/acl-style-files`.

3. Ethayarajh, Duvenaud, and Hirst [6]

4. Ravfogel et al. [13]

## References


[1] Su Lin Blodgett et al. "Language (Technology) Is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485.

[2] Su Lin Blodgett et al. "Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1004–1015. DOI: 10.18653/v1/2021.acl-long.81.

[3] Tolga Bolukbasi et al. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., Dec. 2016, pp. 4356–4364. ISBN: 978-1-5108-3881-9.

[4] Rishi Bommasani and Percy Liang. *Trustworthy Social Bias Measurement*. Dec. 2022. DOI: 10.48550/arXiv.2212.11672.

[5] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases". In: *Science* 356.6334 (Apr. 2017), pp. 183–186. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aal4230.

[6] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. "Understanding Undesirable Word Embedding Associations". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 1696–1705. DOI: 10.18653/v1/P19-1166.

[7] Seraphina Goldfarb-Tarrant et al. "Intrinsic Bias Metrics Do Not Correlate with Application Bias". In: *arXiv:2012.15859 [cs]* (June 2021).

[8] Hila Gonen and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 609–614. DOI: 10.18653/v1/N19-1061.

[9] Abigail Z. Jacobs and Hanna Wallach. "Measurement and Fairness". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 375–385. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445901.

[10] Nikita Nangia et al. "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1953–1967. DOI: `10.18653/v1/2020.emnlp-main.154`.

[11] Aurélie Névéol et al. "French CrowS-Pairs: Extending a Challenge Dataset for Measuring Social Bias in Masked Language Models to a Language Other than English". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8521–8531. DOI: `10.18653/v1/2022.acl-long.583`.

[12] Hadas Orgad and Yonatan Belinkov. "Choose Your Lenses: Flaws in Gender Bias Evaluation". In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Seattle, Washington: Association for Computational Linguistics, July 2022, pp. 151–167. DOI: `10.18653/v1/2022.gebnlp-1.17`.

[13] Shauli Ravfogel et al. "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection". In: *arXiv:2004.07667 [cs]* (Apr. 2020).

[14] Zeerak Talat et al. "You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings". In: *Proceedings of BigScience Episode #5  Workshop on Challenges & Perspectives in Creating Large Language Models*. virtual+Dublin: Association for Computational Linguistics, May 2022, pp. 26–41. DOI: `10.18653/v1/2022.bigscience-1.3`.

[15] Oskar van der Wal et al. *Undesirable Biases in NLP: Averting a Crisis of Measurement*. Nov. 2022. DOI: `10.48550/arXiv.2211.13709`.